VINCA ProData:云计算环境下以数据为中心的应用集成工具

王桂玲 刘晨 张鹏 季光 徐学辉

摘要：在传统服务计算研究工作中往往忽视数据共享和数据流，为了在企业应用集成系统升级时更好地支持数据资源共享和应用协同，本文将着重讨论云计算环境下的数据资源接入、以用户为中心的数据资源集成、数据驱动的业务协同等问题。研究成果可用于跨管理域、基于云资源中心的多信息系统的逻辑集成和信息共享、综合集成类系统特征数据的提取和实时呈现等。

关键词：企业应用集成；数据服务组合；业务流程建模

1 引言

行业或企业传统应用集成（Enterprise Application Integration, EAI）系统在云计算环境下的升级和演化是一个新的挑战，需要从行业信息化的视角，正确利用私有云和行业云的概念。应用集成包含的范围和技术比较广。从集成的对象来说，应用集成技术一般可分为数据集成、流程集成和界面集成这三个层次[1]。其中，又以数据集成和流程集成技术为重点。“业务流程”和“数据”是企业信息系统的核心资产，因此对于现代企业来说，通过业务流程协作进行企业内和企业间的协同、以及通过数据集成实现应用的统一数据视图及查询、分析功能，都对企业的信息化进程发挥着至关重要的作用。从 IBM、Oracle（甲骨文）等各大中间件厂商产品线来看，面向服务的业务流程中间件以及面向服务的数据集成中间件也都是其产品线的核心组成部分。

但是，当前企业应用集成系统面临一些新的挑战。当前企业应用集成系统中的流程集成工具和数据集成工具是割裂的，往往使业务人员对什么时候使用、怎样使用、是使用流程集成工具还是数据集成工具感到困惑[2]。业务人员迫切需要统一的、有机结合的集成工具和编程模型。近两年来，云计算成为一种典型的分布式计算模式，基于云计算的大型分布式应用则呈现出数据密集的特点。例如基于 Map-Reduce[1]进行大规模数据分析的 Google 搜索引擎、基于 Hadoop[2]对大规模企业数据进行检索和分析的系统[3]以及科学工作流系统等。这类应用处理的数据分散在互联网环境下的不同部门和组织，其数据量（或多个并发用户请求涉及的数据量累积达）通常达 TB(10^{12}字节)甚至 PB(10^{15}字节)级，面临的用户也是大规模的，且有些应用的不同用户之间具有通过互联网进行协作的需求，数据共享和数据流问题是这类应用的集成面临的主要矛盾。虽然数据和业务流程在企业应用集成系统中具有同等重要的位置，但传统服务计算研究工作中却往往忽视数据共享和数据流，这使得现有的企业应用集成系统在云计算环境下处理这类数据量大（TB 甚至 PB 级）、数据源格式异构（既有结构化数据和文本数据，也有来自于 Web 的半结构化数据）、数据源动态加入等数据密集型应用的需求时，面临着诸多的挑战。

在上述背景下，本文重点关注下面的问题：

1. 与传统的数据集成相比，在互联网环境下面临的用户是大规模的，且用户的集成需

[1] 谷歌（Google）提出的一个用以支持大型数据集的分布式计算软件框架
[2] 一个支持数据密集分布式应用的免费许可软件框架，由阿帕奇(一个支持开源软件的美国非营利组织)开发

求具有多样性。如何提供一种数据聚合与分析的方法和工具，来满足大规模、多样性的用户集成需求，是一项挑战。为最终用户提供一种简化的手段，使其能够自行进行数据资源的集成，是解决这个问题的一个途径。这种手段被称为“以用户为中心的数据资源集成”。另外，互联网环境下的数据是动态变化的，而传统的数据集成方法中，无论是“物化法”还是“虚拟法”[4]，都难以实时反映数据的变化，或难以支持在运行时动态加入新的数据源。因此，本文下面着重介绍针对以用户为中心进行数据资源集成的两方面相关的技术：一方面是 IT 系统数据资源的服务化技术，解决异构的数据资源如何以统一的服务方式进行封装和抽象。另一方面是以用户为中心的数据服务组合技术，解决数据资源封装为数据服务之后，如何以用户为中心，通过对这些服务的组合完成动态数据资源的集成。

2. 在云计算环境下，企业业务流程相关数据呈现爆炸性产生和积累，如何从这些复杂多样的业务流程和爆炸性增长的数据中，及时有效地获取企业信息系统的整体关键数据，从而把握企业命脉，是一项挑战性的任务。传统工作流和服务组合的做法，往往是以控制流为中心，无论是流程的建模还是执行阶段，只关注流程中的输入/输出等流程执行所必需的数据，而忽视了对系统整体造成影响的数据的建模以及运行时及时获取。例如，传统的业务流程建模方法聚焦于对以下元素进行建模：活动的执行者、执行时间、地点、活动完成的状态等，却忽视了对活动执行后产生的影响、效果，活动执行涉及的信息等元素的显式建模。而这些元素往往才是企业真正关心的。例如，在一个应急场景中，一般情况下，指挥中心难以详细了解千差万别的应急事件处理流程，指挥中心更需要重点掌握当前应急资源的分布、事件处理的结果、与事件处理有关的辅助决策信息等。而这些元素并非活动的输入/输出，在传统的业务流程管理系统中，没有提供对它们事先进行建模的手段，在流程运行时，也无法及时方便地获取这些数据。因此，迫切需要一种新的业务流程建模方法，对活动和活动相关的关键数据进行统一建模，支持在业务流程不断变化的情况下，有效对关键数据进行监控。此外，在流程执行过程中，当监控到企业关键数据发生变化时，也很可能需要根据关键数据的实际情况由人来干预，进行临机决策，临时改进业务流程或者发出警告等。

针对上述问题，本文将着重讨论云计算环境下的数据资源服务化、以用户为中心的数据服务组合、数据驱动的业务流程建模等技术。下面首先对学术界相关研究进行分类探讨，然后以红十字会应急物资管理的一个简化场景为例，探讨我们的工作及其运用。

2 IT 系统数据资源的服务化技术

IT 系统的数据资源具有不同的类型，包括网页、数据库和软件模块等。在这些资源中，将软件模块封装为服务是最为常见的。当前，已经有了很多研究成果和软件工具来帮助完成这一任务。例如，阿帕奇（Apache，详见注 2）组织开发的开源项目 Axis 就能很容易地帮助用户开发基于 Java 对象的 SOAP[3]服务。无论是对网页进行服务化，还是对数据库进行服务化，它们的实质都在于能够根据客户端的请求抽取相应的资源信息，将这些信息转换为预先定义的消息格式（如 SOAP 服务采用的 SOAP 消息），并同时作为响应消息返回给客户端。

对于互联网上的 HTML[4]网页和 Web 数据库资源，由于其服务资源结构化特性差，难以直接将其与其他资源进行组合，因此，需要提供一个手段支持最终用户将 HTML 网页和 Web 数据库服务形式存在的资源封装成易于组合的 XML[5]服务，即“网页信息资源的服务化”。这类服务可以看成一类特定形式的包装器（Wrapper）[5]，用于从网页中抽取出数据并组装

[3] Simple Object Access Protocol 简单对象访问协议
[4] HyperText Markup Language，超文本标记语言
[5] Extensible Markup Language，可扩展标记语言

成 XML。关于包装器的自动与半自动构造，在 Web 信息抽取领域已经开展了大量的研究工作，关于这些工作的详细介绍可在综述文献[6,7]中找到。根据用户是否参与包装器的构造过程，当前的研究工作可以分为用户无监督学习的方法和用户（半）监督学习的方法两类。在我们的工作中，面向的使用者是大量不具备编程知识的普通用户，而现有的包装器构造方法对最终用户的特点以及他们的个性化需求缺少考虑。因此，我们要解决的问题是，如何建立允许用户充分表达其个性化需求的、高效、健壮、高适应的网页信息资源服务化机制，支持最终用户通过简单的“浏览、选定、配置”操作即可完成资源的服务化。

互联网上的资源还以其他形式存在，例如 Open API 和 Web 服务等。这些服务具有不同的接口描述方式或者没有显式的接口描述，服务调用的方式也不同。这些异构性为组合这些不同形态的服务带来了极大的障碍。屏蔽不同服务之间的差异性，为用户提供一个一致的服务模型在某些场合下是必要的，这一过程通常被称为“服务一体化”。针对服务的一体化，工业界已经有多种产品出现。其中最有代表性的是 Oracle 的 AquaLogic 数据服务平台[8]。该平台的主要功能为：将各种异构的数据资源包装为 XML 数据服务。其具体实现机制为：通过考察异构数据源的元数据（如关系数据库的 SQL[6]元数据或 Web 服务的 WSDL[7]），自动生成一到多个物理的数据服务，使数据源的数据能够以 XML 格式表达出来。其中，对于关系数据库，每个表或视图能够生成一个数据服务，其中的每一行数据表达为 XML；对于 Web 服务，每个 Web 服务操作返回类型能够生成一个数据服务，其数据操作由 Web 服务的操作实现；其他数据源类型也可能通过类似的机制形成数据服务。

除了 AquaLogic 数据服务平台，还有 WSO2[8]的 Data Services Server 也支持类似功能，并可将关系数据库，Excel Spreadsheet，Google Spreadsheet 以及 CSV[9]等格式的数据源作为数据服务发布。

尽管已经出现了这些针对数据资源的服务化产品，但是数据服务的标准仍然不统一——虽然各个厂商为自己的产品制订了标准，但是不同厂商的产品所生成的数据服务仍然难以互操作。

3 以用户为中心的数据服务组合技术

以用户为中心的数据服务组合大体可分为三类工作：可视化语言、可视化数据流编程以及电子表格（Spreadsheet）编程。下面，分别对它们进行介绍。

1. 可视化语言

可视化语言类相关工作的特点是：为用户提供一个可视化的操作环境，通过接受用户在该环境中的各种操作（包含对可视化控件的拖拽以及键盘输入），生成操作数据资源的脚本语言，并利用该语言建立数据资源的组合视图。在可视化语言类的工作中，操作数据资源的脚本语言的设计与生成是问题的核心。

这一类相关工作的典型例子包括 AquaLogic、Liquid Data[9]、XQBE[10]、Xing[11]等。AquaLogic 为用户提供了工作区界面。在该界面上，用户可以选择先前通过数据资源服务化生成的源数据服务，同时，构建一个目标数据服务，通过连线设置源数据服务与目标数据服务的对应关系，实现数据映射，并最终生成操作这些数据服务的 XQuery 代码。Liquid Data

[6] Structured Query Language，结构化查询语言
[7] Web Services Description Language，Web 服务描述语言
[8] 一个开源应用软件开发公司
[9] Comma Separated Values，逗号分隔文本文件（亦有译成“逗号分隔”、“逗号分隔型取值”等）

与 AquaLogic 类似，其主要区别在于：Liquid Data 允许数据服务与 XML 文档混合，即可以把静态数据加入进来；但是，如果用户手工编辑了该工具自动生成的 XQuery 代码，则不可再次使用可视化视图，而 AquaLogic 支持双向编辑。

可视化语言类工作的缺点有如下两方面：首先，用户的操作与操作效果是分离开的，用户难以在操作过程中随时看到自己的操作所带来的数据变化；其次，各种可视化语言对数据资源和数据操作的表示缺乏统一的原则，用户对任何一种可视化的表达方式都要进行单独学习。

2. 可视化数据流编程

可视化数据流编程相关工作的特点是：由用户构造数据流，该数据流以数据服务为源头（source），并经过一系列的数据加工与变换，最终聚集到一个数据接收装置（sink）中，通过该数据流的构造实现数据资源的集成。其中，数据流以及数据加工与变换都是以图形化的方式表示的，以便用户理解。

这一类相关工作的典型例子包括 Yahoo Pipes[12]、IBM Damia[13]、VINCA4Science[14]以及 Marmite[15]等。Yahoo Pipes 是一种较有影响力的可视化数据流编程环境。它向用户提供了导入数据、通用数据操作符、字符串操作和数字操作等不同的操作模块，用户在工作区内可以拖拽这些模块，通过管道（pipes）连接这些模块，并在模块中设置数据加工与变换的细节，从而实现数据从 RSS/ATOM/REST 等数据服务流出，经过加工和变换，最终输出给用户的全过程。与 Yahoo Pipes 相比，IBM Damia 是一种类似的工作，其特点在于支持更多的数据服务形式，并且对数据操作模块进行了归类（包括 Augment(扩充)/Merge(合并)/Filter(过滤)/Sort(排序)/ Group(编组)/Union(联合)/Transform(变换)/Publish(发布)等 8 类模块），并将各种多样化的操作向这些有限的操作模块进行集中，以增进用户操作的一致性。

可视化数据流编程有如下两个缺点：首先，与可视化语言类似，用户的操作与操作效果仍然是分离开的，用户难以在操作过程中随时看到自己的操作所带来的数据变化；其次，即使实现较为简单的数据资源整合，如果涉及的数据资源或数据操作较多，会导致数据流非常复杂，影响用户的正常使用。

3. 电子表格（Spreadsheet）编程

电子表格（Spreadsheet）编程相关工作的特点是：为用户提供一个类似于 Excel 电子表格的操作界面，在这个界面里，数据资源以表格形式呈现。用户直接在表格上对数据进行增、删、改等操作，对于比较复杂的数据操作，将其分割为若干可组合、可重复使用的小操作，用户进行每步操作之后能立即观察到该操作带来的数据变化。

这一类相关工作的典型例子包括 SheetMusiq[16]，AMICO[17]，SpreadATOR[18]等。在 SheetMusiq 中，研究人员在二维表上构造了一种电子表格代数，为了在二维表上表达复杂的数据对象，设计了分组（group）代数操作，用于在数据上建立递归的分组关系；通过其他代数操作支持对这种分组数据的查询，并证明其表达能力相当于单块 SQL 语句；另外，还特别支持用户对查询的修改，即通过改变先前的操作实现对数据查询结果的修改。

这一类工作的优点是通过直接在表格上操作数据资源，使得用户能随时查看操作结果，相比于可视化语言和可视化数据流编程，这是一个巨大的进步。另外，由于电子表格是用户熟悉的操作界面，用户不必进行特殊的学习过程就可以实际操作，从而给用户带来了诸多方便。但是该类工作也有一定的局限性：由于电子表格用二维表表示数据，对于有嵌套结构的复杂数据的表示就比较困难，而且通过二维表上的操作建立较为复杂的表达能力容易让用户

难以理解。如图 1(a)所示，这是 SheetMusiq 对一张数据表进行了分组操作后的结果，表上覆盖的背景网格为笔者所加，以便读者理解分组操作的含义。从这个简单的例子可见，如果不辅以其他手段，二维表在表达复杂的数据关系时具有先天的欠缺。

针对二维表的局限性，有人提出了嵌套表与电子表格结合的数据资源集成方法。该方法借鉴了嵌套关系代数中的嵌套表作为数据资源的表示手段，并在嵌套表之上建立各种数据操作，实现来源不同的数据资源的集成。图 1(b)是图 1(a)数据转换为嵌套表所得到的结果。在这个表中可以清晰地看出数据的分组关系。除此之外，相比于二维表，在嵌套表上表达复杂的数据操作（如嵌套查询）也较为容易。嵌套表作为数据资源的表示结构，通过与电子表格的结合，有希望建立起用户友好的、表达能力强的数据资源集成手段。

ID	**Model**	**Price**	**Year**	**Mileage**	**Condition**
872	Jetta	$15,000	2005	50,000	Excellent
901	Jetta	$16,000	2005	40,000	Excellent
304	Jetta	$14,500	2005	76,000	Good
723	Jetta	$17,500	2006	39,000	Excellent
725	Jetta	$18,000	2006	30,000	Excellent
423	Jetta	$17,000	2006	42,000	Good
132	Civic	$13,500	2005	86,000	Good
879	Civic	$15,000	2006	68,000	Good
322	Civic	$16,000	2006	73,000	Good

(a) SheetMusiq 中的一个数据表例子

Model	e0				
	Year	e1			
		Condition	e2		
			ID	Price	Mileage
Jetta	2005	Excellent	872	$15,000	50,000
			901	$16,000	40,000
		Good	304	$14,500	76,000
	2006	Excellent	723	$17,500	39,000
			725	$18,000	30,000
		Good	423	$17,000	42,000
Civic	2005	Good	132	$13,500	86,000
	2006	Good	879	$15,000	68,000
			322	$16,000	73,000

(b)对应的嵌套表格形式

图 1. SheetMusiq 和嵌套表的例子

综合以上对相关工作的调查与分析，我们发现，电子表格提供了较为直观的操作模式。而对于二维表难以表达的数据对象，嵌套表是一种有效的表达形式。因此，电子表格编程，特别是嵌套表与电子表格编程的结合，提供了实现用户为中心的数据服务组合的良好途径。

4 数据驱动的业务流程建模技术

针对数据驱动的业务流程建模，学术界近年来开展了一些研究。传统的流程协作的建模方法，例如范德阿尔斯特（Van der Aalst）[19]的对等网络（P2P）方法能够保证参与组织的“私有工作流”(单个自治机构内部的工作流）很好地满足“公共工作流”(涉及多个自治机构的工作流）的约束，整体性较强，特别适用于边界确定和合作伙伴相对稳定的环境。由于该方法预先定义了流程之间的交互关系，当需求有变动或新的组织需要加入进来时，则首先需要修改“公共工作流”，再根据“公共工作流”来重新对“私有工作流”进行限定。该方法没有消除公共工作流与私有流程之间的依赖关系，无法动态地支持合作伙伴的加入和退出。与上述对等网络方法相反，CrossFlow[20]、DynaFlow[21]等工作普遍采用自底向上的处理方式。处理过程可概括为：首先，将私有工作流中需要对外提供的任务节点发布到一个公共注册库中；然后，在公共注册库中匹配合作伙伴，并形成协作策略以描述合作伙伴在协作中所扮演的角色和职责；最后通过中间件的支持来实现流程之间的相互协作，并对协作的过程进行监控。这种方式的优点是：

- 协作是基于自治机构已有的业务流程，自治机构的自主性更强；
- 自治机构之间可以直接进行交互协作，无须借助于公共流程，灵活性更强，能较好

地适应需求的动态变化。

尽管有这些优点，但是这种方式仍然存在依赖特定流程建模语言、依赖特定的软件系统等不足。随着 Web 服务的出现，万维网联盟（World Wide Web Consortium，W3C）提出了支持对等网络协作的 Web 服务编排定义语言 WS-CDL 1.0[22]。但是该方法只考虑了以活动为中心描述流程之间的依赖，不能支持流程运行过程中实时地监控关键数据。

2005 年，范德阿尔斯特提出了一个支持灵活的业务流程的编程范型 Case handling[23]，这种编程范型是建立在数据对象作为流程基本元素的基础上的。Case Handling 提供数据对象建模。一个数据对象是一组复合信息的抽象。所谓复合信息是指具有一系列不同性质或属性的事物，仅有单个值的事物(例如，宽度)不是数据对象。数据对象通过映射（mapping）建立和外部数据源的映射关系，并通过表格（Forms）来呈现。“Case”可以看作一个流程实例，通过建立数据对象和活动的联系，将数据对象和 Case 关联起来，从而在 Case 运行的过程中，允许查看所有与流程实例关联的数据对象。在数据对象中可以定义事件-条件-动作（Event-Condition-Action，ECA）规则对 Case 进行触发，从而支持不同 Case 之间的协同。Case Handling 支持 Case 运行时的监控，其中包括监控 Case 中活动的执行情况和所有数据对象的信息。其中 Case 的状态由所有数据对象的状态决定。这个工作进一步支持了以数据为中心的流程协作，但是在 Case Handling 中，数据对象的数据源无法接入 Web 上的信息源，缺少数据分析和复杂事件处理的功能，使得流程协作的能力受到限制。

在跨域分布环境下的应用协同往往需要很多子流程进行协作。在运行时，流程结构可能随时改变。如果单纯依赖手工方式进行修改，则可能容易出错，并且由于修改者缺少深刻流程知识而无法发现全部的依赖，导致死锁、延迟，阻碍了整个流程的执行。为此，穆勒（D. Müller）提出了 COREPRO 方法[24]，支持定义复杂的数据结构，并且给出了数据驱动的流程结构自动变换的方法。该方法减少了手工方式改变流程结构的工作量，保证流程之间正确协作。但是他们的方法只适合解决简单流程之间的协作，缺少支持复杂流程类似业务过程执行语言（Business Process Execution Language， BPEL）的过程式描述，并且没有业务活动监控的功能。

以 IBM 的赫尔（R. Hull）[25]为代表的一些学者，进一步意识到结合数据和流程在流程协作中的重要作用，提出了以 Artifact 作为流程协作的基本构建模块。这里 Artifact 是与业务相关的概念实体，典型的 Artifacts 包括购物订单，商品发票，装船单，保险索赔单，客户交互的历史信息等等。他们还给出了以 Artifact 为中心的业务流程建模方法。该方法通过 Artifact 来集成异构业务流程，支持跨域的业务流程。Artifact 需要领域专家建模，它的信息模型对达到业务目标的关键数据进行建模，它的“宏生命周期”（Macro-lifecycle）对公共业务流程的结构（Schema）进行建模。目前有两种生命周期描述方法，一种是基于状态机的方法，另一种是基于声明式的方法。Artifact 和服务的关联有两种方式：一种是通过过程描述进行关联，把服务关联到 Artifact 的状态变迁；另一种是通过声明式的业务规则进行关联。在部署时，引擎根据这种 Artifact 和服务的关联生成描述服务组合结果的业务运行模型（BOM）。在运行时，用户可以实时查看关键性能指标（Key Performance Indicator, KPI）的数据变化。这里也可以在关键性能指标上定义触发规则来改进业务流程或生成警告。这种方法从根本上改变了流程建模的方式，能够解决跨域的流程异构问题，该方法需要领域专家为 Artifact 建模以及为 Artifact 和服务之间建立关联，对领域专家的建模能力要求较高。

上述工作在跨域的流程协作中都给出了各自的方法或架构。从发展的趋势上看，解决方法正从集中式编排发展到分散式编排，从以活动为中心发展到以数据为中心，从预先编排发展到临机编排。在这种趋势下，研究一种分散的、数据驱动的、支持临机编排的业务流程建

模是一件有意义的事情。

5 我们的工作

下面结合应急物资管理领域的一个应用场景说明我们的工作及开发的原型系统。应急物资管理包括应急物资的需求分析、筹措、储存、保障运输、配送和使用直至消耗全过程的管理。应急物资管理的特征包括：突发性、不确定性（持续时间、影响范围、强度大小）、强时效性和全面参与性（例如，不可能指望由哪一个物流中心单独实现全部保障，需要众多的物流中心、物流企业在政府特别机构的组织下来共同参与完成）。下面我们以某地区发生地震后的应急物资管理为场景来详细进行说明。

5.1 场景描述

2010 年某天，某偏远山区发生地震。该地红十字会将负责购买、募集、管理及分发帐篷、方便面和水等救灾物资。应急物资调配是一个复杂的、涉及多部门协同的业务流程。本设计将从中选取一个片段并经过简化，来分析业务需求和可能遇到的问题。该简化的场景描述如下：

场景：灾情发生后，红十字会应急办公室负责指挥整个救灾过程。它首先派遣救援部门赶赴灾区了解灾情，救助灾民，并根据灾情进展上报需要的救灾物资。然后，应急办将通知备灾部门，根据灾情需要出库救灾物资。最后，应急办将协同灾区政府救灾特别机构，派遣足够的运输车辆将救灾物资运抵灾区。

该简化场景中需要协同的部门及其承担的角色如下表所示：

表 1：救灾涉及部门及其职能

部门名称	职能
红十字会应急办公室	负责指挥整个救灾过程，协同其他部门完成救灾流程
红十字会备灾部门	管理红十字会仓库，根据灾情需要出库救灾物资
红十字会救援部门	派遣救援队救助灾民，并根据灾情进展上报需要的救灾物资
灾区政府特别机构	派遣运输车辆将救灾物资运抵灾区

作为整个救灾过程的指挥者，红十字会应急办公室需要及时得到全面、准确的灾情情报和物资需求清单；优化利用有限的物资，发挥物资的最大价值；监控物资流动状态，确保最快将物资合理发放到最需要的灾民手中。为了达成上述目标，应急办需要灵活地协同不同部门的业务流程，来完成及时、合理的调配救灾物资的过程。但是，由于灾情的突发性、多变性、复杂性，协同过程需要应对以下挑战。

- 为了及时作出正确的物资分配决策，应急办需要及时了解关键业务数据的变化情况，而这些关键业务数据很多都是隐藏在其他部门自身业务流程中的。
- 应急办不仅需要获取不同业务流程后隐藏的业务数据，有可能还需要快速汇聚这些数据，对其进行分析，并根据分析的结果进行决策。
- 灾情信息的复杂多变导致应急办难以预先构建一个能满足需求的完整的业务流程。应急办往往需要根据业务数据的变化（如最新灾情信息、物资库存信息、车辆调配信息等）才能决定下一步需要采取的行动。

5.2 ProData：云计算环境下以数据为中心的应用集成工具

为了应对上述挑战，我们开发了一种云环境下以数据为中心的应用集成工具 VINCA ProData。它主要包括两部分：一个是数据集成部分，包括数据对象建模和关键性能指标建模工具、以及数据获取、加工模块；一个是业务流程管理模块，包括业务流程建模工具、业务流程执行引擎以及业务流程监控工具。使用 VINCA ProData 应急集成工具解决 5.1 中所述的实际问题，通常可以分为以下几个关键步骤：

1.　数据对象建模

为了支持红十字会的指挥者在流程运行的每一步过程中，实时地查看关键数据，我们提出了数据对象的概念，并采用嵌套表作为数据对象的基本结构。上述场景可以建立如下三个数据对象：

救援队	地理位置			物资				事件		
	No	经度	纬度	No	类型	需求数量	现有数量	No	经度	纬度

仓库	类型	库存数量	分配情况			
			No	救援队	数量	状态

运输车	地理位置			物资		
	No	经度	纬度	No	类型	数量

图 2.数据对象模型

其中救援队属于红十字会救援部门，仓库属于红十字会备灾部门，运输车属于灾区当地的政府特别机构。由于这些数据对象的数据源可能来自分布在不同部门的数据库或者文件，所以我们需要进行数据的获取、加工。数据对象建模工具主要由各部门的 IT 人员使用，这里采用两边定义-中间汇聚的方式为数据对象建模，其中的两边是指面向业务领域的数据对象和面向 IT 领域的分布的数据源，中间汇聚是指数据的映射和转换。

2.　关键性能指标建模

为了让红十字会指挥者实时看到他们关心的关键性能指标，我们提出了关键性能指标建模，这里关键性能指标是以数据对象为操作数的集合或数字计算公式，在类似电子表格的编程环境中进行计算，并且能够通过图/表的方式呈现。该场景中包含如下关键性能指标：

- 统计(救援队.物资(帐篷、方便面和水).需求数量),
- 统计(救援队.物资(帐篷、方便面和水).现有数量),
- 统计(救援队)

3.　制订业务规则

表 2：业务规则及触发的流程预案

业务规则编号	**业务规则内容**
规则 1	救援队.事件.经度>0&&救援队.事件.纬度>0→一级救援流程预案
规则 2	救援队.事件.经度>0&&救援队.事件.纬度>0→二级救援流程预案
规则 3	救援队.物资(帐篷、方便面和水).需求数量>救援队.物资(帐篷、方便面和水).现有数量→A 库物资出库流程预案

为了能够实时捕捉到关键性能指标的数据变化，支持跨部门的流程之间的动态协作，我们通过业务规则定义事件触发的流程预案。这些业务规则通过基于事件代数的组合算子得

到，并且能够自动转换成事件处理语言 EPL(Esper[10] Processing Language)，实时捕捉关键性能指标的数据变化。表 2 是该场景中的部分业务规则。

4. 编排业务流程

各部门的业务人员编排业务流程，在该场景中包括红十字会救援部门、红十字会备灾部门和灾区政府特别机构三个部门的业务流程。由于灾区政府熟悉当地路况信息，所以运输流程由灾区政府组织的特别机构负责。上述三个部门相对独立，每个部门只处理自己的业务逻辑，在应急物资管理过程中进行无中心控制的动态协作，共同完成应急物资的调配和发放工作。

5. 用户临机决策

当关键性能指标的变化满足业务规则设定的条件时，相应事件被红十字会决策中心捕获，红十字会指挥者结合数据对象的数据，对该事件进行分析，从业务规则的预案中临机决策执行哪一个流程，如图 3 所示。例如，由于救援队到达现场后需要大量物资，红十字会指挥者结合 A 仓库和 B 仓库的库存量，选择执行 A 库物资出库流程预案。并且在调运运输队时，根据各个运输队的位置信息，将物资合理最快地发放到最需要的灾民手中。

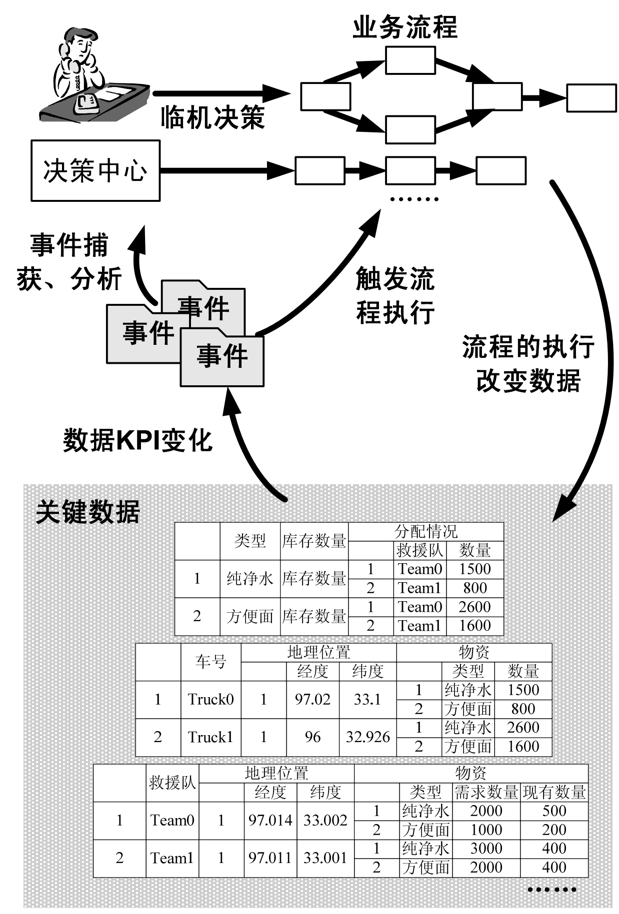


图 3.业务流程的动态协作

6 结束语

近年来，随着云计算的兴起，传统的企业应用集成系统面临升级问题。由于传统服务计算研究工作中却往往忽视数据共享和数据流，使得现有的企业应用集成系统在云计算环境下处理数据密集型应用时，面临着诸多的挑战。本文分析了企业应用集成系统在云计算环境下数据集成以及业务流程建模方面遇到的两大问题，重点对云计算环境下的数据资源服务化、以用户为中心的数据服务组合、数据驱动的业务流程建模等技术进行了分析。最后，以一个应急物资管理场景为例，介绍了一种云计算环境下以数据为中心的应用集成工具----VINCA ProData。在未来的工作中，我们将进一步对云计算环境下以数据为中心的应用集成中数据服务建模、数据服务组合、流程建模及优化等方面的科学问题进行研究。

[10] Esper 是一个事件流处理（Event Stream Processing，ESP）和复杂事件处理（Complex Event Processing，CEP）系统，可以监测事件流并在特定事件发生时触发某些行动

参考文献：

[1] 韩燕波. 王桂玲. 刘晨. 王菁. 赵卓峰., 互联网计算的原理与实践—探索网格、云和 Web X.0 背后的本质问题和关键技术. 科学出版社. 2010.7-

[2] Halevy, A. Y., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A. and Sikka, V. (2005) Enterprise information integration: successes, challenges and controversies. In SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 778-787. ACM.

[3] K.Beyer, V.Ercegovac,R.Krishnamurthy. Towards a Scalable Enterprise Content Analytics Platform. http://sites.computer.org/debull/A09mar/sandeep.pdf

[4] Richard Hull and Gang Zhou, A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches, pp. 481--492, 1996.

[5] J. Hammer, H. Garcia-Molina, S. Nestorov, etc. Template-Based Wrappers in the TSIMMIS System, In Proc. of the International Conference on Management of Data (SIGMOD), Tucson, Arizona, USA, 1997, 532-535

[6] A.H.F. Laender, B.A. Ribeiro-Neto, etc. A brief survey of Web data extraction tools, SIGMOD Record, 2002, 31(2): 84-93

[7] C. Chang, M. Kayed, M. R. Girgis, etc. A Survey of Web Information Extraction Systems, IEEE Trans. Knowl. Data Eng, 2006, 18(10): 1411-1428.

[8] Borkar, Vinayak, Michael Carey, Sebu Koleth, Alex Kotopoulis, Kautul Mehta, Joshua Spiegel, Sachin Thatte, and Till Westmann. "Graphical Xquery in the Aqualogic Data Services Platform." Paper presented at the SIGMOD '10: Proceedings of the 2010 international conference on Management of data 2010.

[9] N. Mangtani and M. Carey. Liquid Data: XQuery-based enterprise information integration. In BEA WebLogic Developer's Journal, April 2003.

[10] Daniele Braga , Alessandro Campi , Stefano Ceri,XQBE(XQuery By Example): A visual interface to the standard XML query language, ACM Transactions on Database Systems (TODS), v.30 n.2, p.398-443, June 2005

[11] Martin Erwig, A Visual Language for XML, Proceedings of the 2000 IEEE International Symposium on Visual Languages (VL'00), p.47, September 10-13, 2000

[12] Yahoo Pipes, Inc. http://pipes.yahoo.com/, 2010.

[13] Simmen, D. E.; Altinel, M.; Markl, V.; Padmanabhan, S.; Singh, A. Damia: data mashups for intranet applications, SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, 2008, pp. 1171-1182.

[14] Jing Wang; Yanbo Han; Shuying Yan; Wanghu Chen; Guang Ji; , "VINCA4Science: A Personal Workflow System for e-Science," Internet Computing in Science and Engineering, 2008. ICICSE '08. International Conference on , vol., no., pp.444-451, 28-29 Jan. 2008

[15] J. Wong and J. I. Hong, Making mashups with marmite: towards end-user programming for the web, In Proc. of the 2007 conference on Human factors in computing systems (CHI), San Jose, California, USA, 2007, pp. 1435-1444.

[16] B. Liu and H. Jagadish. A spreadsheet algebra for a direct data manipulation query interface. In Proceedings of the 35th international conference on Very large data bases. 2009. pp. 417-428.

[17] Ž. Obrenović and D. Gašević. End-User Service Computing: Spreadsheets as a Service Composition Tool. IEEE Transactions on Services Computing. 2008, 1(4). pp. 229-242.

（下转第 7 页）